LOW-LATENCY SOUND-SOURCE-SEPARATION USING NON-NEGATIVE MATRIX FACTORISATION WITH COUPLED ANALYSIS AND SYNTHESIS DICTIONARIES

Tom Barker[⋆], *Tuomas Virtanen*

Tampere University of Technology
Department of Signal Processing
Tampere, Finland
{Thomas.Barker, Tuomas.Virtanen}@tut.fi

Niels Henrik Pontoppidan

Eriksholm Research Centre
Oticon A/S
Denmark
NHP@eriksholm.com

ABSTRACT

For real-time or close to real-time applications, sound source separation can be performed on-line, where new frames of incoming data for a mixture signal are processed as they arrive, at very low delay. We propose an approach which generates the separation filters for short synthesis frames to achieve low latency source separation, based on a compositional model mixture of the audio to be separated. Filter parameters are derived from a longer temporal context than the current processing frame through use of a longer analysis frame. A pair of dictionaries are used, one for analysis and one for reconstruction. With this approach we are able to increase separation performance at low latencies whilst retaining the low-latency provided by the use of short synthesis frames. The proposed data handling scheme and parameters can be adjusted to achieve real-time performance, given sufficient computational power. Low-latency output allows a human listener to use the results of such a separation scheme directly, without a perceptible delay. With the proposed method, separated source-to-distortion ratios (SDRs) can be improved by over 1 dB for latencies below 20 ms, without any affect on latency.

Index Terms— Non-negative matrix factorisation, NMF, source separation, real-time, low-latency.

1. INTRODUCTION

Sound source separation is a topic which has received lots of research effort in previous years, but the majority of separation approaches rely on off-line methods, where the entire audio mixture is available prior to separation. Such approaches are able to make effective use of long temporal context, and can for example, in frame-based approaches, look ahead to glean useful separation information relevant to current frame processing, such as in [1] . In on-line approaches, data must be processed as it becomes available, and subsequent frames can not be used. Such an application, would be for example in real-time source separation for a hearing-aid user, where processing must be performed with the lowest possible latency, typically smaller than 20 ms [2] and even delays of about 3 to 6 ms are detectable [3]. Any noticeable audio-processing delay can cause the effect of received sound being asynchronous with the sound source, which is uncomfortable for the hearing-aid user. Generally, low-latency single-channel speech-enhancement or source separation approaches are based on spectral subtraction or classical Wiener filtering approaches [4], where some estimate of the noise statistics are derived from the noisy signal. Where the sources within a mixture are statistically similar, e.g. in two-talker mixtures, such methods are less effective though.

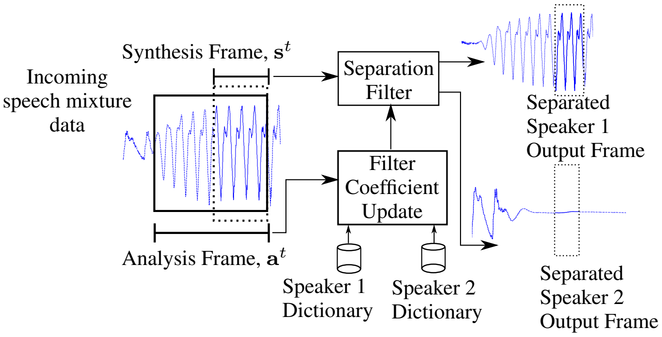
Synthesis Frame, $\mathbf{s}^t$
Incoming speech mixture data
Separation Filter
Filter Coefficient Update
Analysis Frame, $\mathbf{a}^t$
Speaker 1 Dictionary
Speaker 2 Dictionary
Separated Speaker 1 Output Frame
Separated Speaker 2 Output Frame

Fig. 1: Schematic diagram of the proposed source separation approach. As new frames to be processed are formed from incoming data, previous data are retained to provide greater temporal context with which to produce separation filter coefficients.

Compositional model approaches such as exemplar-based non-negative matrix factorisation in [5, 6, 7] represent an audio mixture spectrogram as a summation of non-negative components. In supervised separation approaches, such as proposed in this paper, each sound source present within a mixture is modelled by a non-negative combination of units known as atoms, which are selected from a pre-learned dictionary. These approaches have been shown to be very effective in sound source separation where a well-matched dictionary exists for each source [5], even where sources are of similar types, such as two-talkers. So far though, the effectiveness of such approaches at very low-latency (< 20 ms) has not been well addressed.

Real-time non-negative matrix factorisation based source separation has been considered in [8], where a semi-supervised approach updates the atoms representing noise in the current observation frame from data within a sliding window of variable number of frames with a fixed length of 32 ms. The supervised case is also considered, but frames being separated with this approach are explicitly limited to only their own time context, and is essentially the basic NMF model applied to each observation vector individually. In [9], on-line methods are again considered, with a focus on learning the statistics of unknown sources in the mixture without a-priori information, once again making use of longer analysis frames (64 ms, 75% overlap).

[⋆] The first author performed the work while visiting Eriksholm Research Centre, Denmark.

The research leading to these results has received funding from the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement number 290000 and Academy of Finland grant number 258708.

Real-time separation with pitch-based methods is considered in [10], but pitch estimates are formed over a series of frames, giving a latency greater than 200 ms.This approach is extended specifically for hearing-impaired listeners in [11].

Although low-latency separation using these techniques can be directly implemented by simply using short analysis frames to achieve the desired latency, (as mentioned in [8, 9]), performance is poor when analysis frames are too short and no literature currently exists addressing compositional model based separation with frame lengths below about 20 ms.

With this considered, this paper proposes a low-latency separation approach through parts-based audio representations. The limitation of short analysis frames is overcome by utilising two separate temporal contexts for signal analysis and separated source reconstruction.

By constructing a dictionary for analysis, and coupling each atom within that dictionary to a shorter corresponding reconstruction atom, longer temporal context and hence more information becomes available for modelling each observation vector based on the contributions from two speakers, which improves separation quality. Similar dual-dictionary separation approaches were independently proposed in [12], however this work is more does not explicitly address the low-latency case, which is the focus of this paper.

The remainder of this paper is structured as follows: First, an overview of source separation using compositional models is given, followed by a brief description of the effects of window-length on latency. The proposed model is then described in Section 4, which includes the process by which speaker-specific dictionaries are constructed using such an approach. An evaluation of the effectiveness of the proposed model over the basic single dictionary compositional-model-based separation approaches is presented in Section 5, followed by a summary of the presented work and implementation considerations in Section 6.

2. SOURCE SEPARATION USING DICTIONARY BASED COMPOSITIONAL MODELS

Separation through approximation using linear models has been shown to be effective, in [7]. The spectral magnitude of a mixture is approximated through non-negative summation of components stored within pre-trained dictionaries, with the contributions from each dictionary being used to produce a Wiener filter which is applied to the mixture spectrogram for each source.

A dictionary is defined for each source present within the mixture. Each frame $\mathbf{x}$ to be separated is approximated as a sum of dictionary atoms $\mathbf{d}_k$, weighted by non-negative weights w_k as

$$\mathbf{x} \approx \hat{\mathbf{x}} = \sum_{k=1}^{K} w_k \mathbf{d}_k. \tag{1}$$

An observation vector $\mathbf{x}$ is therefore described by the sum of K components from dictionaries and their respective weights w, where w_k is estimated to minimize a divergence function (typically Kullback-Leibler divergence) between the observation vector, $\mathbf{x}$, and its approximation, $\hat{\mathbf{x}}$. Equation (1) can be rewritten as:

$$\hat{\mathbf{x}} = \mathbf{D}\mathbf{w} \tag{2}$$

where the dictionaries matrix $\mathbf{D}$ is partitioned

$$\mathbf{D} = [\mathbf{D}_1 \mathbf{D}_2] \tag{3}$$

with $\mathbf{D}_1$ and $\mathbf{D}_2$ containing atoms trained on source 1 and source 2 respectively. The weights pertaining to each source are notated $\mathbf{w}_1$ and $\mathbf{w}_2$, and the model can be described as:

$$\hat{\mathbf{x}} = [\mathbf{D}_1 \mathbf{D}_2] \begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{bmatrix}. \tag{4}$$

Table 1: Summary of some of the notations used consistently throughout this paper.

Symbol	Description
$\mathbf{a}^t$	Time-domain analysis frame
$\mathbf{s}^t$	Time-domain synthesis frame
A	Length in samples of $\mathbf{a}_t$
L	Length in samples of $\mathbf{s}_t$
$\mathbf{y}$	Real-valued feature vector formed from $\mathbf{a}_t$
$\mathbf{s}$	Complex-valued synthesis vector formed from $\mathbf{s}_t$
$\mathbf{A}$	Analysis dictionary
$\mathbf{R}$	Reconstruction dictionary
$\mathbf{R}_{:,k}$	The k-th column of dictionary $\mathbf{R}$.
$\mathbf{w}$	Weights vector for a single output frame
$\mathbf{s}_n$	The reconstructed frame for the n-th source in a mixture
n	Subscript referring to the n-th source in dictionaries, weights, or reconstructed frames.

Sources are separated using the above compositional model in the following way. If the complex-valued observation vector to be separated is $\mathbf{y}$, then the separated contribution of the source 1, $\mathbf{s}_1$ is extracted by

$$\mathbf{s}_1 = \mathbf{y} \otimes \frac{\mathbf{D}_1\mathbf{w}_1}{\mathbf{D}_1\mathbf{w}_1 + \mathbf{D}_2\mathbf{w}_2} \tag{5}$$

and similarly for source 2, using the appropriate dictionary and weights in the numerator of Equation 5. The operation can be considered a Wiener filter in the frequency domain, and ensures that reconstructed source estimates sum to the original mixture.

3. DATA PARTITIONING AND LATENCY

For low-latency systems, the time-delay between samples being available for processing and being output as audio should be as low as possible. In frame-based processing schemes, a whole frame of data must be collected and stored before it can be processed for output. We refer to the theoretical minimal delay between a sample incoming into the algorithm and being processed and available for output as ‘algorithmic latency’, T_a, whereas the actual processing time taken can be called ‘computational latency’, T_c. The overall latency T is the sum of these values,

$$T = T_a + T_c. \tag{6}$$

We consider only the constraints of realising low algorithmic latency, since depending on the parameters of a particular processing schema, hardware etc., computation time is non-deterministic.

Since synthesis frames are processed in a block-based manner, a whole frame of input must be captured before the first sample can be output. From a purely algorithmic perspective, sample output can occur as soon as a frame has been processed, regardless of frame overlap. The algorithmic latency of such an approach is therefore the synthesis frame length.

Computational complexity is reduced for non-overlapping frames, but this can result in discontinuities between the last sample of one output frame and the first sample of the next. Greater overlap provides more information which should provide better separation quality than non-overlapping frames, but increases the minimal achievable computational latency.

4. PROPOSED MODEL

In order to maintain low algorithmic latency, processing is applied on short incoming data frames, whilst the filter weights are established

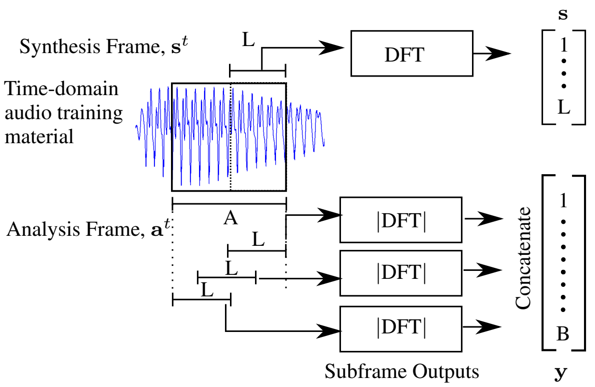

Fig. 2: Feature vector creation. Dictionary atom n for both the analysis and filtering dictionaries are formed from data in the same randomly sampled window.

by examining longer previous temporal context. Since two different frame sizes are used to gather time-domain data for processing, two different atom lengths exist across the coupled dictionaries used in the additive model. For each source, we therefore create separate dictionaries for analysis and reconstruction.

An incoming audio mixture signal is analysed and processed in a frame-based manner, with feature vectors derived from each time-domain frame. Separation is performed by representing feature vectors with a compositional model, where the atoms in each dictionary sum non-negatively to approximate the spectral features of the sources within the mixture. Individual dictionary atoms therefore have the same dimensions as the feature vectors formed from the mixture signal, which are either analysed or filtered in terms of the dictionary contents.

4.1. Frame Lengths and Feature Vectors

For clarity, we now define time domain frame lengths and feature vectors derived from them. Variables are also summarised in Table 1. We refer to the the frame data which is processed for the purposes of separated source reconstruction as the synthesis frame $\mathbf{s}^t$ of length L. A buffer $\mathbf{a}^t$ of previous incoming samples, length A, is maintained (where $A > L$ and A/L is an integer) and referred to as the 'analysis frame'; the temporal context from which the filters to be applied to the processing frame can be derived. The relationships between frames is depicted in Figure 2. Both frames are updated every $L/2$ samples (50% overlap), achieving an algorithmic latency of L whilst reducing computational costs which would be present with higher overlap values.

The analysis feature vector, $\mathbf{y}$, is formed from $\mathbf{a}^t$ by taking the absolute value of the positive frequencies of the discrete Fourier transform (DFT) of analysis sub-frames length L with 50% overlap, and concatenating the resulting $(\frac{2A}{L} - 1)$ subframe outputs into a single vector. The vector effectively describes the magnitude of frequencies present during the past A samples (see Figure 2). The complex-valued frequency-domain synthesis vector $\mathbf{s}$ is formed by taking only the positive frequencies the DFT result of real-valued data in $\mathbf{s}^t$, and so has length $(L/2) + 1$. A Hanning window is applied prior to each DFT on all vectors. $\mathbf{s}$ is filtered at each frame output to produce the separated source estimates as described in Section 4.3.

4.2. Model Dictionaries

For additive model based separation, a dictionary of atoms is typically learned for each speaker in the mixture. We propose the use of coupled dictionaries for each talker, whereby a dictionary of longer analysis atoms is produced alongside a dictionary for source reconstruction.

Explicitly, in 2-talker mixture model, we use one dictionary matrix $\mathbf{A}$ for analysis, and one for source reconstruction, $\mathbf{R}$ where each dictionary comprises talker-specific regions as in Equation 3. The portion of a dictionary trained on source n is notated by the subscript, e.g. $\mathbf{A}_n$, thus:

$$\mathbf{A} = [\mathbf{A}_1\mathbf{A}_2] \tag{7}$$

and

$$\mathbf{R} = [\mathbf{R}_1\mathbf{R}_2]. \tag{8}$$

The k-th atom in the in each dictionary is coupled to the atom at the same index in the alternate dictionary,

$$\mathbf{R}_{:,k} \iff \mathbf{A}_{:,k} \tag{9}$$

by the fact that each was obtained from the same portions of training data, the analysis atoms being derived from a longer previous context than synthesis atoms. The actual dictionary atom creation process is similar to that of feature vector creation depicted in Figure 2. Analysis dictionary atoms are obtained by the same processing as to produce $\mathbf{y}$. Reconstruction dictionary atoms are created similarly to $\mathbf{s}$, except that the real-valued absolute value of the DFT result is stored, as opposed to the complex-valued result present in each $\mathbf{s}$.

Atoms in $\mathbf{A}$ are formed from time domain data of length A whilst L samples are used to form atoms in reconstruction dictionary $\mathbf{R}$. The atoms in $\mathbf{A}$ are used to estimate the weights applied to atoms in $\mathbf{R}$, in order to form the frequency-domain Wiener filters applied to the complex-valued synthesis frame $\mathbf{s}$.

4.3. Analysis and reconstruction using coupled dictionaries

Analysis is performed by learning the weights $\mathbf{w}$ which minimise KL-divergence between analysis vector $\mathbf{y}$ and a weighted sum of atoms from dictionary $\mathbf{A}$ (Equation 10).

$$\underset{\mathbf{w}}{\text{minimize}} \quad f(\mathbf{w}) = \text{KL}(\mathbf{Y}||\mathbf{A}\mathbf{w}) \tag{10}$$

We employ the ASNA algorithm [7, 13] to find the optimal solution due to its rapid computation time and guaranteed convergence, although NMF-based approaches such as in [5] could equally be used, and may offer speed advantages on GPU-based processor architectures.

The learned weights $\mathbf{w}$ are applied to the corresponding coupled dictionary atoms in dictionary $\mathbf{R}$ to form the reconstruction Wiener filters. Filters are applied to the synthesis vector $\mathbf{s}$ at each frame processing step so that the positive frequencies of each frame the of separated source 1, $\mathbf{s}_1$ are reconstructed:

$$\mathbf{s}_1 = \mathbf{s} \otimes \frac{\mathbf{R}_1\mathbf{w}_1}{\mathbf{R}_1\mathbf{w}_1 + \mathbf{R}_2\mathbf{w}_2}. \tag{11}$$

The separated time-domain sources are reconstructed by generating complex conjugates of $\mathbf{s}_n$ and performing the inverse DFT for each frame to be overlap-add reconstructed into a continuous time output.

5. EVALUATION

The separation performance was evaluated computationally, using the source-to-distortion-ratio metric, as defined in [14]. 100 test mixtures were generated, and separated using the proposed approach, with dictionaries being trained on material not part of the test set.

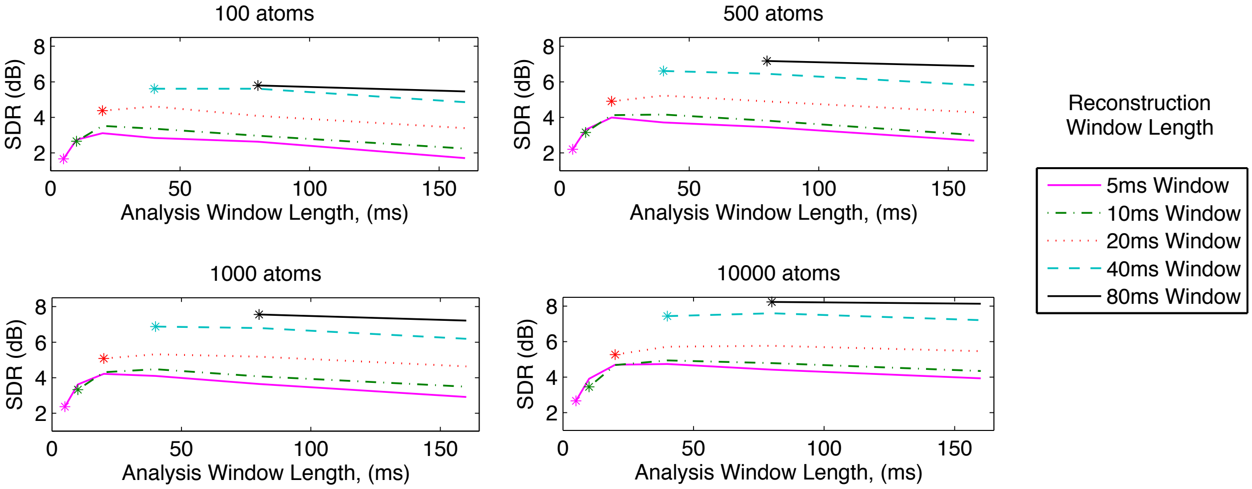

Fig. 3: Separation SDR values for different dictionary sizes. Equal analysis and synthesis frame conditions are marked by $*$.

5.1. Acoustic Test Material

Acoustic test material was taken from the CMU Arctic corpus [15], utterance set B. Pairs of sentences from separate talkers were randomly selected, and summed to form the test mixture. Where one utterance was shorter than the other, the shorter file was padded with zeros prior to summation. All sentences had a samplerate of 16 kHz.

5.2. Dictionary generation

Dictionaries for each speaker and test condition were produced by sampling utterance set A from the CMU Arctic corpus.

For each speaker in the test set, and each test condition (synthesis vs. analysis window length), a dictionary of either 100, 500, 1 000 or 10 000 atoms was used to model the mixture. It has been shown in [7] that larger dictionaries produce greater separation performance since they are able to better model the sources present in a mixture. Larger dictionaries result in greater factorisation times, so shorter dictionaries are worth investigating for time-critical applications. Additionally, as the vector length increases, larger dictionaries are able to better model the greater number of permutations which can be represented by the larger vectors.

5.3. Test Conditions

Each test mixture was separated using the coupled-dictionary-based method described in Section 4.

For each test mixture, a pair of speaker and parameter-specific dictionaries (Section 5.2), are used to model the analysis frames. Analysis buffer lengths of 5, 10, 20, 40, 80 and 160 ms were used with synthesis windows of length 5, 10, 20, 40 and 80 ms. In the case where analysis and reconstruction windows are equal length, the model becomes the standard supervised separation algorithm as described in [8].

Despite being shown to have beneficial effects on separation performance ([1, 13]), sparsity constraints were not applied to the model in these evaluations, although the model could be extended to make use of these. K-means clustering of dictionary atoms from an over-complete initial dictionary as in [7] is also likely to produce performance improvements. Only a single frame overlap value of 50% within the analysis vectors was trialled, although it is possible that the use of greater overlaps, hence longer vectors may produce a better performance at a trade-off against computational latency, and vice versa.

5.4. Experimental Results

Results in Figure 3 show the SDR achieved under various test conditions. The $*$ symbol on each plot line denotes the baseline performance when analysis and synthesis window are of equal length; the basic supervised-separation case. It is seen that an improvement is achieved through use of an analysis frame which is longer than the synthesis frame, where the synthesis frame is 20 ms or below. As a greater number of dictionary atoms is used, this performance gain can also be achieved for 40 ms reconstruction windows. In all cases, using larger dictionaries produces better separation performance than shorter frames, as does using longer reconstruction windows. Where an advantage is gained by use of a longer analysis frame than synthesis frame, the level of improvement reduces as the analysis frame becomes significantly longer than the synthesis frame. For a particular synthesis window length, greatest performance increases are generally achieved when the analysis window is 2-4 times longer.

6. CONCLUSIONS AND DISCUSSIONS

A novel method for increasing source-separation performance in low-latency systems has been proposed. It has been shown that through the use of separate dictionaries for analysis and reconstruction, with atoms derived from different temporal contexts, a significant performance increase is obtained in very low-latency applications ($<$ 20 ms). As larger dictionaries are employed, the maximum performance increases, and it is possible that no improvement is produced for longer synthesis windows since the dictionary is not sufficiently large to effectively describe the possible variation across the resulting longer atoms. Use of a more over-complete dictionary or production of smaller dictionaries through k-means clustering of over-complete data may improve performance further.

In this paper, the proposed algorithm is only evaluated in terms of algorithmic latency. Larger dictionaries and longer vectors increase the required processing, and so would increase computational and overall latency if an efficient processing implementation is not used.

7. REFERENCES

[1] T. Virtanen, "Monaural Sound Source Separation by Nonnegative Matrix Factorization With Temporal Continuity and Sparseness Criteria," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 3, pp. 1066 –1074, march 2007.

[2] S. Laugesen, K.V. Hansen, and J. Hellgren, "Acceptable Delays in Hearing Aids and Implications for Feedback Cancellation," *Journal of the Acoustical Society of America*, vol. 105, no. 2, pp. 1211–1212, 1999.

[3] J. Agnew and J.M. Thornton, "Just Noticeable and Objectionable Group Delays in Digital Hearing Aids," *Journal of the American Academy of Audiology*, vol. 11, pp. 330–336, 2000.

[4] J.M. Kates, *Digital Hearing Aids - Chapter 'Single-Channel Noise Suppression'*, Plural Publishing, 2008.

[5] B. Raj, T. Virtanen, S. Chaudhure, and R. Singh, "Non-negative matrix factorization based compensation of music for automatic speech recognition," in *in proceedings of INTERSPEECH*, 2010.

[6] P. Smaragdis, "Convolutive speech bases and their application to supervised speech separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 15, no. 1, pp. 1–12, 2007.

[7] T. Virtanen, J.F. Gemmeke, and B. Raj, "Active-Set Newton Algorithm for Overcomplete Non-Negative Representations of Audio," *IEEE Transactions on Audio, Speech and Language Processing*, 2013.

[8] C. Joder, F. Weninger, F. Eyben, D. Virette, and B. Schuller, "Real-Time Speech Separation by Semi-supervised Nonnegative Matrix Factorization," in *Latent Variable Analysis and Signal Separation*, vol. 7191 of *Lecture Notes in Computer Science*, pp. 322–329. Springer Berlin Heidelberg, 2012.

[9] Z. Duan, G J. Mysore, and P. Smaragdis, "Online PLCA for Real-time Semi-supervised Source Separation," in *Latent Variable Analysis and Signal Separation*, vol. 7191 of *Lecture Notes in Computer Science*, pp. 34–41. Springer Berlin Heidelberg, 2012.

[10] R. Marxer, J. Janer, and J. Bonada, "Low-Latency instrument separation in polyphonic audio using timbre models," in *Latent Variable Analysis and Signal Separation*, vol. 7191 of *Lecture Notes in Computer Science*, pp. 314–321. Springer Berlin Heidelberg, 2012.

[11] J.H. Gomez, "Low Latency Audio Source Separation for Speech Enhancement in Cochlear Implants," M.S. thesis, Universitat Pompeu Fabra , Barcelona, 2012.

[12] F. Weninger, J. Le Roux, J.R. Hershey, and s. Wantanabe, "Discriminative NMF and its application to single-channel source separation," in *proceedings of INTERSPEECH*, 2014.

[13] T. Virtanen, B. Raj, J.F. Gemmeke, and H. Van hamme, "Active-set Newton Algorithm for Non-Negative Sparse Coding of Audio," in *In Proc. International Conference on Acoustics, Speech, and Signal Processing*, 2014.

[14] E. Vincent, R. Gribonval, and C. Fevotte, "Performance Measurement in Blind Audio Source Separation," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 4, pp. 1462 –1469, July 2006.

[15] J. Kominek and A. W. Black, "CMU Arctic Databases for Speech Synthesis," Tech. Rep., 2003.